BMC Genomic Data

DATA NOTE

Open Access

Draft genome of *Roscoea Debilis*, the first genome in the alpine ginger *Roscoea* (Zingiberaceae)

Xiao-Chang Peng[1†], Ao-Dan Huang[1†], Wen-Jing Wang[1], Gui-Sheng Xiang[2], Li Li[1,3*] and Jian-Li Zhao[1*]

Abstract

Objectives *Roscoea* is a Sino-Himalayan alpine genus in pantropical family Zingiberaeae. As traditional Tibetan medicinal plants, many species of this genus are threatened by digging, logging, land clearance, grazing and climate change. *Roscoea debilis* is an endemic species in the Hengduan Mountains with a narrow distribution range. In this study, the assembled and annotated genome of *Roscoea* was presented in order to furnish significant resources for comparative and functional genomic investigations. The first complete reference genome of *Roscoea* is expected to shed light on research on conservation and evolutionary biology.

Data description A chromosome-level genome of 1601.04 Mb was obtained for *R. debilis* by combining Illumina short reads (107.28 Gb) and PacBio Hi-Fi reads (64.08 Gb), achieving high-quality sequencing coverage of roughly 67 × and 40 ×. The assembly was additionally assisted by 271.65 Gb Hi-C data (169 ×), which resulted in a contig N50 of 136.17 Mb and a scaffold N50 of 90.48 Mb. Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment results revealed that most of the core embryophyta genes (98.7%) in the BUSCO dataset (embryophyta_odb10) were successfully identified. Additionally, 96.44% of the genomic sequences were accurately mapped onto twelve pseudochromosomes.

Keywords *Roscoea*, Hi-Fi Genome, Gene prediction, Genome annotation

Objective

Roscoea is an alpine genus of the Sino-Himalayan distribution, characterized by a unique pattern of disjunctive distribution from the Hengduan Mountains to the Himalayas, and is known for its orchid-like flowers and diverse mating systems [1–3]. *Roscoea* plants have been traditional Tibetan medicines since ancient times in their place of origin [4–6], and also exhibit the potential to be used for horticultural and ornamental purposes [7]. In addition, *Roscoea* plants have a variety of mating systems, including specialized outcrossing, facultative outcrossing, and early selfing [8–10]. They are an ideal group for exploring the evolution of mating systems and their response to environmental changes in mountainous

[†] Xiao-Chang Peng and Ao-Dan Huang have equal contributions to this research.

*Correspondence:
Li Li
lili0426@ynu.edu.cn
Jian-Li Zhao
jianli.zhao@ynu.edu.cn
[1] Ministry of Education Key Laboratory for Transboundary Ecosecurity of Southwest China, Yunnan Key Laboratory of Plant Reproductive Adaptation and Evolutionary Ecology, Institute of Biodiversity, School of Ecology and Environmental Science, Yunnan University, Kunming, Yunnan 650504, China
[2] College of Agriculture and Biotechnology, Yunnan Agricultural University, Kunming, Yunnan 650500, China
[3] Chongqing Key Laboratory for Utilization and Evaluation of Special Chinese Materia Medica Resources, Chongqing Academy of Cinsese Materia Medica, Nanan, Chongqing 400065, China

regions. Coupled with their high-altitude distribution characteristics, they deserve more attention in both practical application and ecological significance. However, many species of this genus are threatened by digging, logging, land clearance, grazing and climate change.

A high-quality reference genome is useful for various aspects of plant ecology and biology research; however, there is currently no complete genome material available for *Roscoea*. *Roscoea debilis* is an endemic species in the Hengduan Mountains, with a narrow distribution range and a unique mating system of early selfing [11]. In this study, the high-quality genome of *R. debilis* was assembled and annotated using a combination of Illumina sequencing and PacBio Hi-Fi sequencing. To the authors' knowledge, this is the first complete reference genome of *Roscoea*, which can provide important basic data for future research on *Roscoea.*

Data description

Leaf samples of *R. debilis* were obtained from an individual located in Tonghai, Yunnan, Southwest China (102.69272°N, 24.073225°E). The genome DNA was extracted from the leaves, and the total RNA was extracted from all tissues, to form four sequencing libraries are formed. Subsequently, Hi-Fi long-read whole genome sequencing (WGS) was performed through PacBio Sequel II sequencers, and short-read WGS, RNA-seq, and Hi-C sequencing were carried out with Illumina NovaSeq 6000 sequencers. Under the MGI platform, a 150 bp paired-end mode was applied for both short WGS and RNAseq. The cross-linked and lysed cells for Hi-C library were digested using Dnp II restriction enzyme. Hi-Fi sequencing generated approximately 64.08 Gb of long-read WGS data (data file 1). Illumina produced about 107.28 Gb of short-read WGS data (data file 2). RNA-seq produced around 6.15 Gb of data (data file 3) for auxiliary commentary. Hi-C sequencing generated approximately 271.65 Gb of data (data file 4).

Following the process of sequencing, the genome survey of *R. debilis* was conducted using BBmap [12] with Illumina short-read data. The findings suggested that the *R. debilis* genome was diploid and highly heterozygous, with a heterozygosity rate of 0.7%, and an estimated genome size of around 1.65Gb. The genome was subsequently assembled using Hi-Fiasm [13] with clean Hi-Fi long reads. Hi-Fi reads were assembled using Purge_dups [14] to eliminate redundant sequences. The assembled sequence was further polished with Illumina short reads and NextPolish [15], yielding a preliminary assembled genome version (contig level). In order to achieve a genome at the chromosomal level, Hi-C reads were aligned to major contigs using chromap [16]. Pair reads that were valid were adopted for the subsequent assembly. YaHS [17] was employed for Hi-C scaffolding. The resulting genome contigs achieved a combined length of 1601.04 Mb, which closely matched the estimated size. The contig N50 was 90.48 Mb, and 96.44% of the contig sequences were aligned to 12 pseudochromosomes. The

Table 1 Overview of all data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Raw long whole genome Hi-Fi sequencing reads 1	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA015916 https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112174 [25]
Data file 2	Raw long whole genome Hi-Fi sequencing reads 2	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA015916 https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112175 [26]
Data file 3	Raw short whole genome Illumina sequencing reads	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA015916 https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112173 [27]
Data file 4	Raw RNA-seq reads	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA015916 https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1136897 [28]
Data file 5	Raw Hi-C reads	Fasta file (.fastq)	CNCB Big sub-Genome Sequence Archive (GSA) Accession number CRA015916 https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112176 [29]
Data file 6	Assembled genome	Fasta file (.fa)	CNCB Big sub-Genome Warehouse (GWH) Accession number GWHES-OX00000000 https://ngdc.cncb.ac.cn/gwh/Assembly/84773/show [30]
Data file 7	BUSCO assessment of the assembly using Embryphyta database	7z file (.7z)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]
Data file 8	Hi-C clustering heat map	pdf file (.pdf)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]
Data file 9	Predicted gene	Gff3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]
Data file 10	Predicted gene - CDS	Fasta file (.fas)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]
Data file 11	Predicted gene - Protien	Fasta file (.fas)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]
Data file 12	Predicted repetitive sequences	Gff file (.gff)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]
Data file 13	Genome feature statistics	7z file (.7z)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]
Data file 14	Gene annotation using GO, KO, interPro, Pfam, KEGG, SwissProt, Meryl, and SwissProt databases	7z file (.7z)	Figshare, https://doi.org/10.6084/m9.figshare.25712205.v2 [31]

total scaffolds length was 1544.01 Mb, and the scaffold N50 was 136.17 Mb. The Benchmark Universal Single-Copy Orthologous (BUSCO) [18] evaluation, based on the embryophyte_odb10 dataset, demonstrated that 98.7% of genes were successfully identified, indicating a high level of integrity in the assembly.

RepeatModeler [19], RepeatMasker [20], and RepBase+Dfam databases were utilized for homologous and de novo prediction of repeat sequences. A total of 1,307,034,346 bp of repeats were predicted, accounting for 84.65% of the whole genome, with 878,082,544 bp being LTR.

For gene structure prediction, Hisat2 [21] was adopted for RNA-seq based prediction, miniprot [22] was used for homology prediction, and augustus [23] was employed for de novo prediction. Eventually, EVM [24] was integrated to obtain the final annotation results. A total of 30,745 genes were predicted.

Limitations

Despite the high integrity of the draft genome of *R. debilis* through contig N50, scaffold N50, and BUSCO results, there are still 28 gaps in the assembly of this version. Therefore, elaborate research on *Roscoea* requires more complete genome assembly. In the future, the *R. debilis* genome can be assembled to T2T (Telomere-to-Telomere) level using either ONT (Oxford Nanopore Technologies) sequencing or deeper Hi-Fi sequencing, which will yield a greater amount of genomic information.

Acknowledgements
The authors appreciated BerryGenomics for providing high-quality sequencing.

Author contributions
P. X-C and H. A-D were jointly responsible for collecting samples, generating and analyzing sequencing data, and finally wrote the manuscript. L. L assisted in assembling the genome. W. W-J and X. G-X assisted in genome annotation. P. X-C, H. A-D and Z. J-L conceived and designed this project. All of the authors have read and approved the final version of this manuscript.

Funding
The study was supported by the National Natural Science Foundation of China (no. 42171057 and 41871047).

Data availability
The raw sequence data (including Illumine, Hi-Fi，Hi-C and RNA-seq data) reported in this paper have been deposited in the Genome Sequence Archive (GSA) in National Genomics Data Center, China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number CRA015916. The assembled genome reported in this paper has been deposited in the Genome Warehouse (GWH) in National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation, under accession number GWHESOX00000000. The results of genome annotations have been uploaded to Figshare. Please see Table 1 for details and links to the data.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 May 2024 / Accepted: 20 August 2024
Published online: 27 August 2024

References

1. Paudel BR, Shrestha M, Burd M, Li Q-J. Dual mechanisms of autonomous selfing in *Roscoea nepalensis* (Zingiberaceae). Ecology. 2021;102(7). https://doi.org/10.1002/ecy.3337.
2. Cowley EJ. The genus Roscoea: Royal Botanic Garden, Kew, UK; 2007.
3. Zhao J-L, Xia Y-M, Cannon CH, Kress WJ, Li Q-J. Evolutionary diversification of alpine ginger reflects the early uplift of the Himalayan-Tibetan Plateau and rapid extrusion of Indochina. Gondwana Res. 2016;32:232–41. https://doi.org/10.1016/j.gr.2015.02.004.
4. Rawat S, Jugran AK, Bhatt ID, Rawal RS. Influence of the growth phenophases on the phenolic composition and anti-oxidant properties of *Roscoea procera* Wall. in western Himalaya. Journal of Food Science and Technology-Mysore. 2018;55(2):578 – 85. https://doi.org/10.1007/s13197-017-2967-z
5. Srivastava S, Misra A, Kumar D, Srivastava A, Sood A, Rawat AKS. Reversed-phase high-performance Liquid Chromatography ultraviolet Photodiode Array Detector Validated Simultaneous Quantification of six Bioactive Phenolic Acids in *Roscoea purpurea* Tubers and their *In vitro* Cytotoxic Potential against Various Cell Lines. Pharmacognosy Magazine. 2015;11(44):S488-S95. https://doi.org/10.4103/0973-1296.168944
6. Luo M, Wan H, Lin H. Species and distribution of *Roscoea* in China and their Medicinal uses. Chin Wild Plant Resour. 2008;27(5):35–741.
7. Misra A, Srivastava S, Verma S, Rawat AKS. Nutritional evaluation, antioxidant studies and quantification of poly phenolics, in *Roscoea purpurea* tubers. BMC Res Notes. 2015;8:324. https://doi.org/10.1186/s13104-015-1290-x.
8. Zhang Z-Q, Li Q-J. Autonomous selfing provides reproductive assurance in an alpine ginger *Roscoea Schneideriana* (Zingiberaceae). Ann Botany. 2008;102(4):531–8. https://doi.org/10.1093/aob/mcn136.
9. Zhao J-L, Gugger PF, Xia Y-M, Li Q-J. Ecological divergence of two closely related *Roscoea* species associated with late quaternary climate change. J Biogeogr. 2016;43(10):1990–2001. https://doi.org/10.1111/jbi.12809.
10. Paudel BR, Shrestha M, Burd M, Adhikari S, Sun Y-S, Li Q-J. Coevolutionary elaboration of pollination-related traits in an alpine ginger (*Roscoea purpurea*) and a tabanid fly in the Nepalese Himalayas. New Phytol. 2016;211(4):1402–11. https://doi.org/10.1111/nph.13974.
11. Fan YL, Li QJ. Stigmatic fluid aids self-pollination in *Roscoea Debilis* (Zingiberaceae): a new delayed selfing mechanism. Ann Botany. 2012;110(5):969–75. https://doi.org/10.1093/aob/mcs169.
12. BBMap_39.06. https://sourceforge.net/projects/bbmap/. Accessed 23 November 2023.
13. Cheng HY, Concepcion GT, Feng XW, Zhang HW, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–. https://doi.org/10.1038/s41592-020-01056-5.
14. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36(9):2896–8. https://doi.org/10.1093/bioinformatics/btaa025.
15. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36(7):2253–5. https://doi.org/10.1093/bioinformatics/btz891.
16. Zhang H, Song L, Wang X, Cheng H, Wang C, Meyer CA, et al. Fast alignment and preprocessing of chromatin profiles with Chromap. Nat Commun. 2021;12(1). https://doi.org/10.1038/s41467-021-26865-w.
17. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39(1). https://doi.org/10.1093/bioinformatics/btac808.
18. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.
19. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.

20. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protocols Bioinf. 2009;Chap 4:4101–44. https://doi.org/10.1002/0471250953.bi0410s25.
21. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–. https://doi.org/10.1038/s41587-019-0201-4.
22. Li H. Protein-to-genome alignment with miniprot. Bioinformatics. 2023;39(1). https://doi.org/10.1093/bioinformatics/btad014.
23. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9. https://doi.org/10.1093/nar/gkl200.
24. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9(1):r7. https://doi.org/10.1186/gb-2008-9-1-r7.
25. Peng X-C. Draft Genome of *Roscoea debilis*-Hi-Fi reads 1. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2024. https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112174
26. Peng X-C. Draft Genome of *Roscoea debilis*-Hi-Fi reads 2. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2024. https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112175
27. Peng X-C. Draft Genome of *Roscoea debilis*-Illumina reads. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2024. https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112173
28. Peng X-C. Draft Genome of *Roscoea debilis*-RNA-seq reads. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2024. https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1136897
29. Peng X-C. Draft Genome of *Roscoea debilis*-Hi-C reads. China National Center for Bioinformation (CNCB)-Genome Sequence Archive (GSA). 2024. https://ngdc.cncb.ac.cn/gsa/browse/CRA015916/CRR1112176
30. Peng X-C. Draft Genome of Roscoea debilis. China National Center for Bioinformation (CNCB)- Genome Warehouse (GWH). 2024. Accession number GWHESOX00000000 https://ngdc.cncb.ac.cn/gwh/Assembly/84773/show
31. Peng X-C. Draft Genome of *Roscoea debilis*. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25712205.v2

Publisher's note